3E

УДК 004.89, 004.934

***Т.В. Ермоленко*[1], *М.А. Панфилова*[2]**
[1]Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк
[2]Государственный университет информатики и искусственного интеллекта, г. Донецк, Украина
etv@iai.donetsk.ua, mirpanph@yandex.ru

Влияние GSM-сжатия на идентификационные акустические признаки, характеризующие речевой поток в целом

Статья посвящена исследованию влияния алгоритма сжатия GSM 6.10, используемого современной сотовой связью, на характеристики речевого сигнала. В работе описаны интегральные признаки речевого сигнала, используемые в современных системах идентификации диктора, выполнено численное исследование устойчивости значений этих признаков к сжатию с потерями.

Введение

В последнее время очень активно развиваются разнообразные технические системы определения индивидуальности говорящего по речевым характеристикам. Потребность в разработке таких систем диктуется наличием широкого круга различных приложений, где требуется подтвердить или опознать определенную личность. Особенно это актуально при разработке систем криминалистической экспертизы звукозаписей.

Известно множество цифровых методов выделения идентификационных признаков из речевого сигнала [1-4]. Однако из-за низкого качества исследуемых фонограмм при проведении фоноскопических экспертиз возможно использование лишь тех акустических признаков, которые являются инвариантными к влиянию искажений фонограмм. К ним относятся признаки, описывающие статистические характеристики амплитудно-частотной спектральной плотности речевого сигнала и основного тона сопоставимых по контексту фрагментов речи. Эти признаки получили название интегральных, поскольку, в отличие от признаков «тонкой спектральной структуры» сопоставимых звуков, измеренных синхронно с моментом возбуждения голосовых связок, они вычисляются на участках речевого сигнала, содержащих сопоставимые слова и фразы, длительностью около 10 секунд. Интегральные признаки характеризуют речевой поток в целом и определяют групповую принадлежность говорящего.

При вычислении спектральных признаков речевой сигнал подвергается фильтрации «гребенкой» полосовых цифровых фильтров и последующему спектральному анализу.

Все известные на сегодняшний день аппаратно-программные комплексы фоноскопических экспертиз, представленные на рынке СНГ, при формировании акустических идентификационных признаков используют методику системы «Диалект», разработанную по заказу ФСБ России в 1995 году. В этой системе для получения интегральных признаков используется гребенка из 21 цифрового фильтра в диапа-

зоне частот от 0 до 3662 Гц, ширина полосы пропускания каждого из фильтров составляет 174 Гц [4]. Таким образом, предложенная разработчиками «Диалекта» методика не дает возможности выбирать количество полосовых фильтров и настраивать границы их полос пропускания.

В криминалистической практике зачастую возникает ситуация, когда образцы речи, по которым формируют эталоны, и сигнал, подлежащий идентификации, записаны в разных условиях. Последний, в большинстве случаев, записан в условиях использования телефонной, преимущественно сотовой, связи, в которой активно используется цифровой формат, получивший название GSM [5], [6]. Звук этого формата получается кодированием специальным алгоритмом сжатия, в результате чего происходит потеря полезного сигнала (объем данных о речевом сигнале сокращен примерно в 5 – 10 раз), которая влияет на распознавание говорящего. Таким образом, потребность в проверке устойчивости признаков идентификации дикторов по их речи возрастает с каждым днем, делая эту задачу актуальной.

Цель работы – исследование влияния алгоритма сжатия GSM на идентификационные признаки, характеризующие поток речи в целом. Для достижения цели поставлены и решены следующие задачи:

– сделать обзор методов формирования интегральных акустических признаков, используемых современными системами идентификации говорящего;

– программно реализовать описанные методы с учетом психоакустических принципов восприятия и провести численное исследование влияния алгоритма GSM-сжатия на значения полученных признаков.

Обзор современных методов вычисления интегральных признаков для автоматической идентификации личности по голосу

В качестве интегральных спектральных признаков используются следующие наборы идентификационных признаков:

– нормированные значения энергетического спектра;
– нормированные средние значения энергетического спектра;
– относительное время пребывания сигнала в полосах энергетического спектра;
– нормированное время пребывания сигнала в полосах энергетического спектра;
– медианные значения энергетического спектра речи в полосах;
– относительная мощность спектра речи в полосах;
– величины вариации огибающих энергетического спектра речи;
– нормированные величины вариации огибающих энергетического спектра речи;
– значения коэффициентов кросскорреляции спектральных огибающих между полосами энергетического спектра;
– значения компонент гистограммы распределения частоты основного тона.

Для получения наборов интегральных признаков сигнал разбивается на окна постоянной длины. Для анализа речевого сигнала длина окна выбирается с учетом периода основного тона (ОТ) и составляет около 20 мс, поскольку в нормальной речи параметры возбуждения не изменяются быстро. На каждом из окон вычисляется кратковременный энергетический спектр с помощью фильтрации гребенкой цифровых фильтров.

В качестве полос пропускания в рамках данной работы была использована барк-шкала [7], связанная с критическими полосами слуха, а также темперированная

музыкальная шкала [8]. Такой выбор обусловлен психоакустическими принципами восприятия [7]. В табл. 1 приведены значения граничных частот 25 критических диапазонов для барк-шкалы, по формуле (1) получают значения центральных частот полос по темперированной музыкальной шкале, которые зависят от номера ноты N (самой низкой ноте «Ля» с целым значением частоты 55Гц соответствует номер $N = 33$). Представление локальных значений мощности в пределах психоакустических шкал позволяет моделировать процесс обработки речевого сигнала человеческим ухом.

$$F(N) = 440 \cdot 2^{\frac{(N-69)}{12}} = 55 \cdot 2^{\frac{(N-33)}{12}}. \quad (1)$$

Таблица 1 – Центральные частоты и границы полос по барк-шкале

№ полосы	Границы полос, Гц	№ полосы	Границы полос, Гц	№ полосы	Границы полос, Гц
1	0 – 100	9	920 – 1080	15	4400 – 5300
2	100 – 200	10	1080 – 1270	16	5300 – 6400
3	200 – 300	11	1270 – 1480	17	6400 – 7700
4	300 – 400	12	1480 – 1720	18	7700 – 9500
5	400 – 510	13	1720 – 2000	13	9500 – 12000
6	510 – 630	18	3700 – 4400	24	12000 – 15500
7	630 – 770	13	1720 – 2000	25	15500 – …
8	770 – 920	14	2000 – 2320		

После фильтрации гребенкой из M цифровых фильтров (в зависимости от использованной шкалы) речевой сигнал может быть представлен в виде двумерного массива значений кратковременных энергетических спектров (спектральных срезов), полученных на каждом окне анализа:

$$\{x(i,j)\}_{i,j=1}^{M,J}, \quad (2)$$

где $x(i,j)$ – значение энергии сигнала на выходе i-го полосового фильтра в j-м спектральном срезе; J – общее количество окон на анализируемом отрезке сигнала.

Введем операцию нормировки массива $\{a(i)\}_{i=1}^{N}$ по $2k+1$ точкам:

$$d(a(i),k) = \left\{ \begin{array}{l} d(a(k+1),k), i = \overline{1,k} \\ \dfrac{a(i)}{\sum\limits_{j=i-k}^{i+k} a(j)}, i = \overline{k+1, N-k} \\ d(a(N-k),k), i = \overline{N-k+1, N} \end{array} \right\}. \quad (3)$$

Признаки первой группы – нормированные значения энергетического спектра:

$$X(i) = \frac{x(i)}{\sum\limits_{i=1}^{M} x(i)}, \quad (4)$$

где $x(i)$ – среднее по строке массива:

$$x(i) = \frac{1}{J}\sum_{j=1}^{J} x(i,j). \quad (5)$$

Признаки второй группы – нормированные согласно (3) значения энергетического спектра (4):

$$X_H(i) = d(X(i),3). \quad (6)$$

Нормировка значений признаков (4) вводится для снижения их зависимости от линейных (частотных) искажений речевого сигнала при прохождении его по тракту звукозаписи.

В третьей группе признаков – относительное время пребывания сигнала в полосах энергетического спектра, значение каждого i-го признака вычисляется по формуле (7):

$$t(i) = \frac{\Delta J(i)}{J}, \quad (7)$$

где $\Delta J(i)$ – количество спектральных срезов, при которых энергия в i-й полосе превышает среднее значение (5).

В четвертой группе признаков – нормированное время пребывания сигнала в полосах энергетического спектра, полученное согласно (8):

$$t_H(i) = \frac{t(i)}{\sum_{i=1}^{M} t(i)}. \quad (8)$$

Пятую группу составляют признаки нормированных медианных значений энергетического спектра, вычисляемые по формуле (9):

$$m_H(i) = \frac{m(i)}{\sum_{i=1}^{M} m(i)}, \quad (9)$$

где $m(i)$ – медианное значение энергетического спектра для полосы i-го фильтра.

Признаки шестой группы – нормированные значения мощности спектра в полосах, которые вычисляются согласно (10):

$$P_H(i) = d(P(i),3), P(i) = m(i)/\Delta J(i). \quad (10)$$

Нормировка мощности спектра вводится по той же причине, что и нормировка средних значений энергетического спектра (6), для снижения влияния линейных искажений в трактах передачи сигнала.

Седьмую группу признаков составляют вариации огибающих энергетического спектра:

$$V(i) = \nu(i) \cdot x(i),\ \nu(i) = \frac{1}{J}\sum_{j=1}^{J}\left(x(i,j) - x(i)\right)^2. \quad (11)$$

Признаки восьмой группы – нормированные значения вариаций, которые вычисляются согласно (12):

$$V_H(i) = d(V(i),3). \quad (12)$$

Нормировка каждой i-й компоненты вариации огибающей спектра необходима для снижения влияния частотных искажений на значения признаков.

Девятую группу признаков составляют коэффициенты кросскорреляции $R(i,k)$, которые вычисляются по формуле (13):

$$R(i,k) = \frac{(J-1)\sum_{j=1}^{J}\{x(i,j) - x(i)\}\cdot\{x(k,j) - x(k)\}}{J\sqrt{\sum_{j=1}^{J}\{x(i,j) - x(i)\}^2}\cdot\sqrt{\sum_{j=1}^{J}\{x(k,j) - x(k)\}^2}}, i \neq k. \quad (13)$$

Десятую группу интегральных акустических признаков составляют признаки ОТ речи, а именно значения компонент гистограммы распределения частоты основного тона (ЧОТ). Эти признаки предназначены для описания особенностей распределения значений ОТ речи говорящего в диапазоне от 50 до 400 Гц.

По результатам анализа исследовались следующие компоненты гистограммы распределения ЧОТ: средняя частота; максимальная частота; минимальная частота; асимметрия плотности распределения; эксцесс плотности распределения; график распределения плотности.

Для определения величины ЧОТ по речевому потоку применяется следующий алгоритм. Из анализируемого отрезка речи с помощью экспериментально установленных порогов устраняются фрагменты, соответствующие низкоэнергетичным элементам речи, и участки, имеющие высокую частоту пересечения нулевого уровня сигнала (в основном, согласные звуки). Полученный таким образом сигнал разбивается на окна длиной около 20 мс. Величина ЧОТ определяется на каждом из окон с помощью кепстрального анализа. Метод оценивания ОТ на основе кепстрального анализа сводится к отысканию пика в области возможных значений ОТ, координата пика дает оценку периода ОТ.

Первые шесть групп интегральных признаков отражают своеобразие формы спектра голосовых импульсов у различных лиц и особенности фильтрующих функций их речевых трактов.

Признаки вариаций огибающих энергетического спектра (11) и нормированных вариаций огибающих энергетического спектра (12) характеризуют особенности речевого потока, связанные с динамикой перестройки артикуляционных органов речи говорящего.

Коэффициенты кросскорреляции (13) являются интегральными характеристиками речевого потока, отражающими своеобразие взаимосвязи или синхронности движения артикуляционных органов речи говорящего.

Группа интегральных признаков ОТ характеризует индивидуальность статистических распределений значений ЧОТ речи говорящего, которая, в свою очередь, является параметром колебаний голосовых связок и определяет, главным образом, групповую принадлежность голоса человека.

Численное исследование

Для проведения численного исследования, предназначенного для изучения влияния алгоритма GSM-сжатия на интегральные идентификационные характеристики речевого сигнала, были записаны речевые фрагменты, принадлежащие 10 дикторам (мужчинам и женщинам с разными голосовыми данными). Для каждого из дикторов сделаны 10 записей длительностью не менее 5 секунд. Диктор наговаривал набор из 11 слов, которые не содержат невокализованных звуков. Образцы речи для обеспечения максимального приближения их характеристик к исходным аналоговым сигналам записывались в формате WAV PCM с частотой дискретизации 22050, глубиной битности 16 бит. Запись осуществлялась в монорежиме. Кроме того, для каждого из дикторов была сделана одиннадцатая запись в формате GSM 6.10 WAV с частотой дискретизации 8 кГц. Все записи были созданы с помощью программы Audacity 1.3.12-beta.

Идентификационный анализ компрессированных речевых реализаций 10 дикторов и образцов их речи, зафиксированных в формате WAV PCM, проводился по вышеописанным 10 группам интегральных признаков с использованием трех типов гребенок фильтров: по барк- и темперированной музыкальной шкале, а также по гре-

бенке фильтров, предложенной разработчиками системы «Диалект». Для чего было создано специальное программное обеспечение, реализующее вычисление значений компонент гистограммы распределения ЧОТ и групп признаков (4), (6) – (13), с возможностью настройки границ полос пропускания используемых фильтров и визуального анализа полученных признаков каждой группы в виде графиков их значений.

Допустимые значения вариативности (внутридикторская вариативность) интегральных признаков рассчитывались на основе статистических оценок, полученных по образцам речи, в качестве оценки междикторской вариативности использовалась дисперсия средних значений исследуемых признаков.

Результаты проведенного идентификационного анализа приведены в табл. 2.

Таблица 2 – Результаты идентификационного анализа компрессированных речевых реализаций

Группа признаков	Вероятность принадлежности речевой реализации одному лицу		
	барк-шкала	Темперированная музыкальная шкала	«Диалект»
нормированные значения энергетического спектра	0,93	0,96	0,9
нормированные средние значения энергетического спектра	0,93	0,96	0,9
относительное время пребывания сигнала в полосах энергетического спектра	0,34	0,4	0,27
нормированное время пребывания сигнала в полосах энергетического спектра	0,31	0,35	0,25
медианные значения энергетического спектра речи в полосах	0,95	0,97	0,91
относительная мощность спектра речи в полосах	0,94	0,95	0,91
величины вариации огибающих энергетического спектра речи	0,91	0,92	0,87
нормированные величины вариации огибающих энергетического спектра речи	0,96	0,97	0,95
значения коэффициентов кросскорреляции спектральных огибающих между полосами энергетического спектра	0,91	0,95	0,89
значения компонент гистограммы распределения ЧОТ	0,95		

На рис. 1 показаны гистограмма распределения ЧОТ, полученная для образца речи, записанного в формате WAV PCM, на рис. 2 – для речевой реализации того же диктора, записанной в формате GSM 6.10.

Анализ гистограмм распределения ЧОТ показал, что степень совпадения их характеристик для сигналов после компрессии и соответствующих образцов речи высокая (различия не более полутона), т.е. значения компонент гистограммы распределения ЧОТ, полученные по сжатому речевому сигналу, изменяются в пределах внутридикторской вариативности.

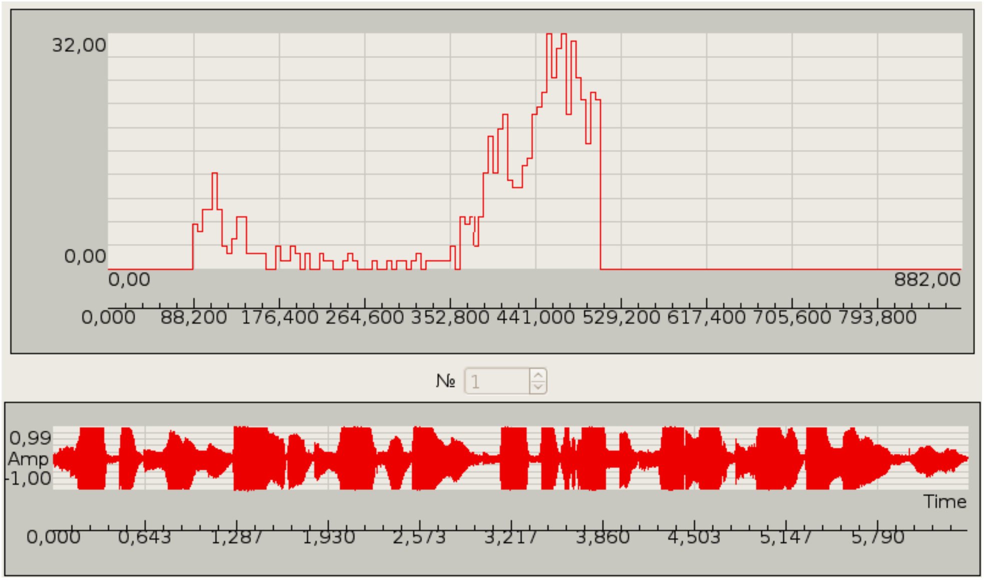


Рисунок 1 – Гистограмма распределения ЧОТ (вверху), полученная по речевому материалу диктора Ж1, записанного в формате WAV PCM (внизу – график амплитудно-временного представления соответствующего речевого сигнала)

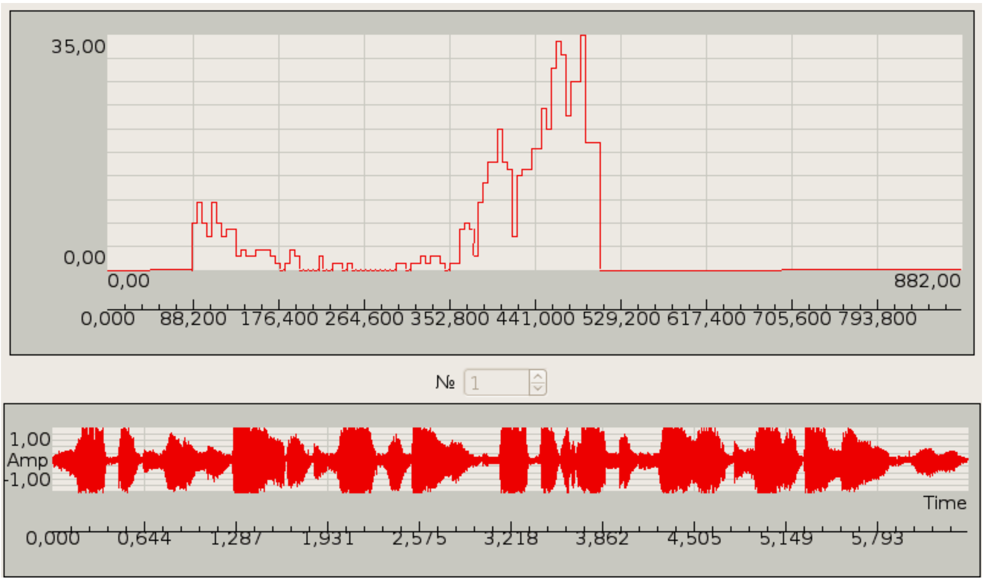


Рисунок 2 – Гистограмма распределения ЧОТ (вверху), полученная по речевому материалу диктора Ж1, записанного в формате GSM 6.10 WAV (внизу – график амплитудно-временного представления соответствующего речевого сигнала)

Выводы

Анализируя полученные результаты, можно сделать вывод о перспективности использования темперированной музыкальной шкалы для задания границ полос пропускания фильтров, на основе которых вычисляются спектральные признаки, характеризующие речевой поток в целом.

Высокая степень совпадения характеристик гистограмм распределения ЧОТ и спектральных признаков (4), (6), (9) – (10), (12) – (13), отражающих характеристики

спектральной плотности речевого сигнала, полученных для сигналов после компрессии, с характеристиками, полученными по соответствующим образцам речи, говорит об устойчивости этих групп признаков к GSM-сжатию.

Разница значений вариаций огибающих энергетического спектра (11) в области высоких частот превышает порог внутридикторской вариативности, что делает эту группу признаков непригодной для идентификационного исследования компрессированных сигналов. Группы признаков (7) – (8) – относительное и нормированное время пребывания сигнала в полосах энергетического спектра – показали свою полную непригодность для идентификационных исследований.

Система связи GSM развилась в глобальный стандарт второго поколения, занимающий лидирующие позиции в мире, на основании чего можно полагать, что цифровые фонограммы этого формата все чаще будут попадать в сферу уголовного и гражданского судопроизводства в качестве доказательной базы. Однако специфика такого формата вводит новые проблемы в проведение фоноскопических экспертиз, поскольку сигнал подвергается интенсивному кодированию с удалением существенной порции криминалистически значимой информации о речи абонента. Таким образом, возникает необходимость проверки робастности существующих методик идентификации личности по голосу и разработки новых экспертных методик исследования цифровых фонограмм, что определяет практическую значимость данной работы.

Литература

1. Женило В.Р. Компьютерная фоноскопия / Женило В.Р. – М. : Академия МВД России, 1995. – 208 с.
2. Каганов А.Ш. Криминалистическая экспертиза звукозаписей / Каганов А.Ш. – М. : Юрлитинформ, 2005. – 272 с.
3. Идентификация дикторов на основе сравнения статистик основного тона голоса / С.Л. Коваль, П.В. Лабутин, Е.В. Малая, Е.А. Прощина // Сб. трудов XV международной научной конференции «Информатизация и информационная безопасность правоохранительных органов». – М. : Академия управления МВД России. – 2006. – С. 324-327.
4. Идентификация лиц по фонограммам русской речи на автоматизированной системе «Диалект» / под ред. Фесенко А.В., Попов Н.М., Линьков А.Н., Кураченкова Н.В., Байчаров Н.В. В/ч 34435. – 1996. – 102 с.
5. Иванов И.Л. Экспертное исследование формата GSM [Электронный ресурс] / Иванов И.Л. – Режим доступа : http://www.illidiy.orel.ru/Pub/publ6.htm
6. Тимко Е.В. Проблемы криминалистического исследования цифровых фонограмм [Электронный ресурс] / Е.В. Тимко, К.Ю. Усков // Труды Киевского НИИ судебных экспертиз. – 2001. – Режим доступа : http://www.expert.com.ua
7. A Review of Algorithms for Perceptual Coding of Digital Audio Signals / Ted Painter, Andreas Spanias // Proc. International Conf. on Digital Signal Processing (DSP). – Santorini (Greece), 1997. – P. 179-205.
8. Шилов Г.Е. Простая гамма (устройство музыкальной шкалы) / Шилов Г.Е. – М. : Физматгиз, 1963. – 20 с., ил.

Т.В. Єрмоленко, М.О. Панфілова

Вплив GSM-стиснення на ідентифікаційні акустичні ознаки, що характеризують мовний потік у цілому
Стаття присвячена дослідженню впливу алгоритму стиснення GSM 6.10, що використовується сучасним стільниковим зв’язком, на характеристики мовного сигналу. У роботі описано інтегральні ознаки мовного сигналу, що використовуються в сучасних системах ідентифікації диктора, виконано чисельне дослідження стійкості значень цих ознак до стиснення із втратами.

T.V. Yermolenko, M.A. Panfilova

Influence of GSM-compression on Identification Acoustic Features that Characterize a Speech Flow as a Whole
The article is devoted to research of influence of GSM 6.10 compression algorithm, which is employed within modern cellular networks, on speech signal features. The technique for integral speech signal features computation which is often used in modern speaker identification systems is described. Also computational investigation results of these features robustness to compression with losses are shown.

Статья поступила в редакцию 16.07.2010.